基于蚁群聚类的动态加权 PPI 网络复合物挖掘 *

胡 健[a]，朱海湾[b]，毛伊敏[b]

(江西理工大学 a.应用科学学院 信息工程系; b.信息工程学院，江西 赣州 341000)

摘 要：针对基于蚁群聚类的蛋白质复合物挖掘算法中，静态 PPI 网络难以真实反映细胞的动态特性，收敛速度较慢、聚类准确性和召回率不高等问题进行了研究，提出一种基于模糊粒度和紧密度的蚁群聚类的动态加权 PPI 网络复合物挖掘方法（joint fuzzy granular and closeness degree ant colony clustering-DPC，FGCDACC-DPC）。首先基于动态 PPI 网络的拓扑特性和生物特性设计了综合性权值度量（comprehensive weight metric，CWM），准确描述了蛋白质之间的相互作用；其次根据复合物的基本特征，构建一组稠密且高度共表达的复合核，然后设计模糊粒度和紧密度的拾起放下模型对其余节点聚类，降低了计算复杂度和随机性，加快聚类速度；最后基于功能信息传递和时序功能相关的思想分别构建了局部和全局权值更新策略，实现不同代蚁群和不同时刻网络之间的功能信息传递，提高聚类准确性。将 FGCDACC-DPC 算法应用在 DIP 数据上进行复合物挖掘，实验结果表明该算法的精度和召回率较高，能够较准确地识别蛋白质复合物。

关键词：蚁群聚类；模糊粒度；动态 PPI 网络；功能传递；蛋白质复合物

中图分类号：TP399 doi: 10.19734/j.issn.1001-3695.2018.07.0518

Mining protein complexes based on ant colony clustering in dynamic weighted PPI networks

Hu Jian[a], Zhu Haiwan[b], Mao Yimin[b]

(1. *Dept. of Information Engineering, College of Applied Science, b. School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou Jiangxi* 343100, *China*)

Abstract: Since static PPI networks are difficult to truly reflect the dynamic character of cells, the convergence speed is slow, cluster precision and recall is low in mining protein complex based on ant colony clustering, this paper proposes an ant colony clustering algorithm based on fuzzy granular and closeness degree to mine protein complexes in dynamic weighted PPI network, named FGCDACC-DPC. First, based on the topological and biological characteristics of the PPI network, a comprehensive weight metric (CWM) is designed to accurately describe the interaction between proteins. Second, this method constructs a series of dense and highly co-expressed complex core based on the basic characteristic of the complexes, then it employs the picking and dropping operations, which based on fuzzy granular and closeness degree, to cluster the nodes in PPI networks, in order to reduce effectively the computational complexity and randomness, speed up the clustering speed. Finally, this algorithm designs a local and global strategy founded on function transmission and timing functional relevance theory for weight's update, which achieve the function transmission between different generations of ant colonies and networks at different times to effectively improve clustering accuracy. FGCDACC-DPC algorithm is used to mine protein complexes on DIP data. Experimental results demonstrate that this algorithm has better performance on precision and recall, which is more reasonable to identify protein complexes.

Key words: ant colony clustering; fuzzy granular; dynamic protein-protein interaction network; function transmission; protein complex

0 引言

蛋白质是维持一切生命活动的基础，其功能一般是通过与蛋白质之间的相互作用表现出来的。一个生命体内，由若干蛋白质复合物之间的相互作用构成的网络叫做蛋白质交互网络，而蛋白质复合物又是在同一空间和同一时间下共同完成某项功能的蛋白质集合。由于蛋白质复合物挖掘不仅能帮助人们理解生命活动的过程、预测功能未知的蛋白质，还为疾病诊断和药物研制提供理论基础[1]，因此蛋白质复合物挖掘成为现如今的一研究热点。但目前大多识别复合物的聚类算法都是基于静态 PPI 网络，由于这类算法不能较真实地反映蛋白质相互作用网络的动态变化[2]，因此基于动态 PPI 网络挖掘蛋白质复合物的研究显得尤为重要。

随着 PPI 数据和蛋白质序列数据的日益完善，不少学者逐渐转向基于计算的复合物挖掘的研究，也提出了许多传统的挖掘算法，如有基于密度的 MCODE 算法[3]，基于划分的 RNSC 算法[4]和基于层次的 Jerarca 算法[5]等。但这些算法都存在一定的不足，有的算法对于稀疏网络效果不佳，有的算法检测不到重叠的复合物，有的算法对噪声敏感等等。近年来，研究人员又提出一些新的复合物检测方法，如基于流模拟的检测方法[6]、基于核心-附件结构的检测方法[7]、谱聚类算法[8]以及群智能算法[9-12]等。而功能流算法的聚类结果受给

收稿日期：2018-07-11；**修回日期：**2018-08-27 **基金项目：**国家自然科学基金资助项目（41562019，41530640）；江西省自然基金资助项目（20161BAB203093，GJJ161566）；江西省教育厅科技项目（GJJ151528GJJ151531）；省社科规划项目（13YD020）

作者简介：胡健（1967-），男，江西赣州人，教授，博士，主要研究方向为数据挖掘、软件工程；朱海湾（1995-），硕士研究生，主要研究方向为数据挖掘（1411870893@qq.com）；毛伊敏（1970-），女，教授，博士，主要研究方向为数据挖掘、地理信息系统等.

定参数的影响较大，基于核心—附属结构的聚类方法复杂度较高，不适用于大规模 PPI 网络，谱聚类算法在数据降维后又回到传统聚类方法上。群智能优化算法具有强大的全局寻优能力，并且具有较强的鲁棒性。尤其是蚁群算法具有和其他群智能算法相比独特的优势，蚁群算法本身就能直接聚类，不需要借助其他聚类算法，能够充分发挥群智能算法的优势。目前蚁群算法已成功应用于 PPI 网络复合物和功能模块挖掘，成为该领域一个新的研究热点。刘志军[9]提出一种蚁群优化的 PPI 网络功能模块检测算法 NACO-FMD，该方法设计一种更有目的性的函数指导蚁群寻优，得到较好的聚类效果。刘红欣[10]提出一种蚁群聚类的功能模块检测算法 ACC-FMD，该方法通过拾起放下模型对节点聚类，以最优解更新相似度函数，通过不断迭代使聚类结果趋于最优，最后对聚类结果合并过滤。这些蚁群聚类算法在应用于大规模 PPI 网络上都需要进行大量的拾起放下，以及合并过滤等操作，导致收敛速度慢，求解时间过长。吕嘉伟等人[11]提出一种基于多粒度模型的蚁群优化算法 MGRACO-FMD，试图提升收敛速度，但聚类结果准确性不高。Lei 等人[12]提出一种基于连接强度的 PPI 网络蚁群优化聚类算法，该算法时间开销有所降低，但查全率也较低。这些算法在提升时间性能的同时，正确率和查全率都有所降低。以上算法都将蛋白质相互作用网络视为静止不变的，但是静态 PPI 网络不能真实反映细胞内部的动态变化[13]，进而影响蛋白质复合物挖掘的准确性，因此基于动态 PPI 网络挖掘蛋白质复合物更为合理。目前许多学者展开了从动态 PPI 网络中挖掘蛋白质复合物方面的研究。Tang 等人[14]利用基因表达数据和静态 PPI 网络，构建一个规定统一阈值的时序蛋白质相互作用（time course protein interaction networks，TC-PIN），并且将其成功应用于蛋白质功能模块挖掘。由于各个蛋白的基因表达水平不一致，设置统一阈值会导致构建的 PPI 网络不准确，进而影响聚类效果。Hu 等人[15]取消统一阈值，将各个蛋白质的平均表达水平作为评判该蛋白是否为活性的标准，结合复合物信息和结构域信息构建动态加权网络，并提出蛋白质功能预测方法 D-PIN，实验表明该方法具有较高的准确率，但召回率相对较低。Su 等人[16]提出一种基于动态加权 PPI 网络复合物挖掘算法 GECluster，该方法首先利用 GO-Slim 对动态网络加权，其次根据种子节点扩充的策略挖掘蛋白质复合物。该方法只采用基因本体信息度量蛋白质之间的功能相似性，并未融合多种数据，因此不能很好地反映蛋白质之间的相互作用。Yi 等人[17]利用边聚集系数和持续共表达长度对各个蛋白质加权，提出一种基于核附属的蛋白质复合物检测方法 DCA，该算法的加权方式融入了复合物演化的时序特征，能够较好地描述蛋白质之间的相似性。同年，Zhao 等人[18]利用复合物的时序功能保持特征，结合蚁群聚类，提出一种新的复合物识别算法。该算法从一种新的视角去分析复合物的挖掘方法，而不仅仅只在聚类方法上进行创新。该方法的聚类精确度较高，但是算法的召回率一般，可能与权值度量及蚁群搜索方式有关。虽然基于动态 PPI 的蛋白质复合物挖掘取得了一定的成效，但如何有效利用基因表达谱过滤假阳性数据，如何合理整合 PPI 数据和多元生物信息，并提出有效的加权方式来减少构建的网络与真实网络之间的差距，仍需深入研究。此外蚁群算法应用于大规模 PPI 网络聚类问题中，需进行大量拾起放下和过滤操作，导致收敛速度慢，同时由于算法随机性较大，导致准确率和召回率普遍不高，这些问题仍亟待解决。

针对以上问题，本文首先利用静态网络和基因表达数据，将大规模静态 PPI 网络划分为多个小规模瞬态 PPI 网络，有效降低 PPI 网络中假阳性对复合物检测结果的影响，并且在一定程度上解决蚁群聚类算法应用于大规模 PPI 网络收敛速度慢的问题。基于动态网络，本文主要做了以下几个方面的工作：a）结合点边聚集系数、GO 功能相似性和基于皮尔逊相关系数得到的表达谱的相似性，提出一种综合性权值度量方法，对网络加权并添加新的相互作用，有效降低假阴性，进而提高召回率;b）基于动态加权网络，提出一种基于蚁群聚类的复合物检测方法 FGCDACC-DPC,首先根据复合物的生物特性，构建一组功能相似、小而稠密的复合核，蚁群随机选择一个作为初始位置,其次采用模糊粒度相似性函数对其余节点进行聚类，聚类完成之后根据紧密度舍弃与复合物连接不紧密的节点，优化聚类过程;c）考虑到蚁群之间的信息传递和 PPI 网络的时序相关性，提出一种局部和全局权值更新策略，通过在不同代蚁群和相邻时刻网络之间不断传递最优解信息使聚类结果趋于最优，并且能够有效减少访问却不被拾起的次数，进而加快聚类速度、提高聚类准确性。实验结果表明，该算法的聚类效果较好。

1 动态加权 PPI 网络构建

1.1 动态 PPI 网络构建

目前动态网络已引起人们广泛的关注，动态网络的构建是一个根据基因表达谱数据对静态网络不断调整优化的过程，动态网络的定义如下所示。

定义 1 动态网络[19] $DG=\{G_1,G_2...G_i...G_t\}$，$G_i=\{V_i,E_i\}$ 是 i 时刻的瞬态网络，t 表示时刻数，$V_i=\{v_{i1},v_{i2},...,v_{in}\}$ 表示 i 时刻下表达的蛋白质集合，$E_i=\{e_{i1},e_{i2},...,e_{im}\}$ 表示 i 时刻下蛋白质相互作用集合。

动态 PPI 网络是基于基因表达谱数据构建的。基因表达数据对揭示蛋白质和 PPI 网络的动态属性极其重要，包含每个蛋白质的 36 个时刻的基因表达水平（3 个周期，每个周期 12 个时刻）。本文首先采用文献[15]中的公式，将 36 个时刻合并为 12 个时刻，分别取三个周期对应时刻的平均值作为该时刻的基因表达值，计算公式如下：

$$T_u(i)=\frac{T_u(i)+T_u(i+12)+T_u(i+24)}{3} \tag{1}$$

基于这 12 个时刻的基因表达值，通过蛋白质的共表达性来构建动态 PPI 网络。由于不同蛋白质基因表达水平不同，不适合采用统一阈值判断其活性[20]。因此在本文中，如果某蛋白质在该时刻下的基因表达水平大于自身的平均表达水平，则认为该蛋白质在该时刻下表达。每个蛋白质在 12 个时刻下的平均基因表达值如下所示：

$$T_u^{'}=\frac{\sum_{i=1}^{12}T_u(i)}{12} \tag{2}$$

其中：$T_u(i)$ 表示蛋白质 u 在 i 时刻下的基因表达值，$T_u^{'}$ 表示蛋白质 u 的平均基因表达值。

本文整合基因表达数据，对静态 PPI 网络不断调整，进而构建动态 PPI 网络。基本思想如下：如果蛋白质 u,v 在 PPI 网络上存在相互作用，且在同一时刻表达，那么本文认为蛋白质 u,v 在该时刻下的瞬态网络确实存在边，否则认为蛋白质 u,v 之间的边是虚假的而剔除；如果蛋白质 u,v 在 PPI 网络中不存在相互作用，但在同一时刻表达，考虑到 PPI 网络中存在假阴性，本文根据 1.2.4 节中 GO 功能相似性和皮尔逊相关系数的取值大小，决定是否在它们之间新增一条边。

1.2　动态 PPI 网络加权

针对 PPI 网络中存在大量假阴性数据的问题，本文基于 PPI 网络的拓扑特性，结合 GO 功能相似性和共表达的皮尔逊相关系数，对动态网络加权，并且添加新的相互作用，能在一定程度上增加蛋白质相互作用的可信度。

1.2.1 PPI 网络拓扑特性

边聚集系数[21]是网络拓扑特性中最重要的一种，不仅考虑了边在网络中的重要程度，还能评估节点 u,v 邻居之间的紧密程度。边聚集系数的定义如下：

$$E_{cc}(u,v)=\frac{\tan_{u,v}}{\min(d_u-1,d_v-1)} \tag{3}$$

其中：$\tan_{u,v}$ 表示节点 u,v 共同构成三角形的个数。d_u,d_v 分别表示节点 u,v 的度。

由于边聚集系数只考虑边的重要性，没有考虑节点的重要性，把两个节点的权值都看做是 1。因此，引入能够反映节点聚集程度的点聚集系数[22] 对边聚集系数加以改进，提出一种融合节点和边的双重拓扑特性的点边聚集系数 CE_{cc}。点边聚集系数公式如下：

$$CE_{cc}(u,v)=\frac{\tan_{u,v}\times C_u\times C_v}{\min(d_u,d_v)} \tag{4}$$

其中：$\tan_{u,v}$ 和 d_u,d_v 如式（3）所示，C_u,C_v 表示节点 u,v 的点聚集系数，计算公式如下：

$$C_v=\frac{2N_v}{d_v(d_v-1)} \tag{5}$$

其中：d_v 表示节点 v 的度，N_v 表示由节点 v 的邻居节点之间组成的边数目。

节点 u 所有关联边的点边聚集系数之和定义如下：

$$SoCE_{cc}(u)=\sum_{v\in Neigh(u)} CE_{cc}(u,v) \tag{6}$$

例如：如图 1 所示，在该网络中有 9 个节点，19 条相互作用。根据式（4）计算每一条边的点边聚集系数，再使用式（6）计算 $SoCE_{cc}$ 值来评价该节点的重要程度。计算过程如下。

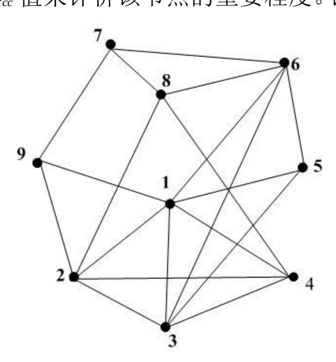


图 1　ppi 网络示意图

Fig. 1　An example of protein-protein interaction network

首先，使用式（5）计算各个节点的点聚集系数。

$$C_1=0.4667,C_2=0.5,C_3=0.6,C_4=0.6667,C_5=1,$$
$$C_6=0.4,C_7=0.3333,C_8=0.3333,C_9=0.3333$$

其次，使用式（6）计算每个节点的重要程度。

$$SoCE_{cc}(1)=\sum_{v\in Neigh(u)} CE_{cc}(1,v)=CE_{cc}(1,2)+CE_{cc}(1,3)+\ldots$$
$$\ldots+CE_{cc}(1,4)+CE_{cc}(1,5)+CE_{cc}(1,6)+CE_{cc}(1,9)=0.957246$$

同理可得 $SoCE_{cc}(2)=\sum_{v\in Neigh(u)} CE_{cc}(2,v)=0.60722$，可知节点 1 的重要性大于节点 2。

1.2.2　蛋白质 GO 功能相似性

由于生物实验的局限性，PPI 网络中往往存在大量噪声数据，如果仅以网络拓扑特性衡量两个蛋白质之间的相互作用程度，比较片面。因此本文引入 GO 功能注释信息能够有效降低假阴性带来的负面影响，提高网络的可靠程度。研究表明，两个蛋白质的 GO 注释语句的交集越多，功能就越相似，则出现在同一复合物的概率越大。受文献[23]的启发，本文将两个蛋白质 u,v 的 GO 功能相似性公式定义如下：

$$FS(u,v)=\frac{|f_u\cap f_v|^2}{|f_u||f_v|} \tag{7}$$

其中：$|f_u\cap f_v|$ 表示蛋白质 u,v 共同的 GO 术语数目，$|f_u|,|f_v|$ 分别表示蛋白质 u,v 的 GO 术语数目。

1.2.3　基因共表达的皮尔逊相关系数

引入皮尔逊相关系数来度量两个蛋白质共表达的强弱程度，能够在一定程度上抑制 GO 注释信息的引入带来的假阳性的升高。蛋白质 u,v 的皮尔逊相关系数定义如下：

$$Pcc(u,v)=\frac{1}{k-1}\sum_{i=1}^{t}\left(\frac{E_{xp}(u,i)-\bar{E}_{up}(u)}{\sigma(u)}\right)\left(\frac{E_{xp}(v,i)-\bar{E}_{xp}(v)}{\sigma(v)}\right) \tag{8}$$

其中：k 为样本数，i 为在基因表达数据中的时刻数，$E_{xp}(u,i),E_{xp}(v,i)$ 分别表示蛋白质 u,v 在 i 时刻下的表达值，$\bar{E}_{up}(u),\bar{E}_{up}(v)$ 和 $\sigma(u),\sigma(v)$ 表示在所有时刻下的平均表达值和标准方差，$Pcc(u,v)\in[-1,1]$。

1.2.4　动态网络加权

基于动态网络的网络拓扑特性，整合 GO 注释信息以及基因表达数据对网络进行加权，该加权策略体现出一致性和互补性。一致性表现在能够共同反映蛋白质相互作用的可信度，权值越大可信度越高。互补性表现在，由于引入的 GO 注释信息中可能包含虚假信息，会导致假阳性升高，进而导 FS 值有所升高，负的 Pcc 值能够在一定程度上降低影响；由于 PPI 数据的假阴性和 GO 注释信息的不完整性，会导致边的权值有所下降，正的 Pcc 值能够在一定程度上弥补。

对于动态 PPI 网络 $DG=(V,E)$ 中任意两个蛋白质 u,v 之间存在相互作用，则它们之间的相互作用权值计算公式如下：

$$CWM(u,v)=CE_{cc}(u,v)+FS(u,v)+Pcc(u,v) \tag{9}$$

否则，蛋白质 u,v 之间的相互作用权值如下：

$$CWM(u,v)=FS(u,v)+Pcc(u,v) \tag{10}$$

其中：$E_i^{'}=\{(u,v)\mid u,v\in V_i,(u,v)\notin E_i,CWM_i(u,v)>0\}$ 表示通过 GO 功能注释信息和皮尔逊相关系数新增的边，如果 $W_i(u,v)<0$，则把蛋白质 u,v 之间的权值当做 0，即 $W_i(u,v)=0$。加权后的动态网络定义如下。

定义 2　动态加权网络 $DWG=\{G_1,G_2,\ldots,G_i,\ldots G_t\}$，$G_i=\{V_i,E_i\cap E_i^{'},CWM_i\}$ 是 i 时刻的加权网络，t 表示时刻数，$V_i=\{v_{i1},v_{i2},\ldots,v_{in}\}$ 表示 i 时刻下表达的蛋白质集合，$E_i=\{e_{i1},e_{i2},\ldots,e_{im}\}$ 表示 i 时刻下蛋白质相互作用集合，$CWM_i=\{cwm_{i1},cwm_{i2},\ldots,cwm_{im}\}$ 是权值的集合。

1.3　动态加权 PPI 网络的构建过程

构建动态加权 PPI 网络，能够减少假阳性和假阴性数据，使得网络真实可靠。具体构建过程如下所示:

输入：静态 PPI 网络，基因表达谱数据，GO 注释信息

输出：各个时刻下的动态加权网络 DWG_i

a)根据式（1）将 36 个时刻合并为 12 个时刻，然后根据式（2）将表达值低于平均值的蛋白质过滤。

b)构建动态网络。在某个瞬态子网下，如果蛋白质 u,v 在静态 PPI 网络上存在相互作用且共表达，则在该时刻网络中添加一组相互作用；如果蛋白质 u,v 在静态网络上不存在相互作用但共表达，则判断式（10）是否大于 0，大于 0 则添

加一组相互作用，否则不添加。

c)分别根据式（9）和（10）对 12 个动态子网进行加权。

2　算法描述

2.1　蚁群聚类算法

基于拾起放下规则的蚁群聚类算法[24]是由 Lumer 和 Faiela 提出的，其主要思想是：将数据散落在一个二维平面上，随机生成部分蚂蚁，蚂蚁有两种状态：负载和空载。蚂蚁在负载状态时，将负载数据与可见范围内的数据进行相似度对比，若和周围数据相似，则将该数据放下，否则继续随机移动；蚂蚁若为空载状态，判断该位置的数据和周围数据的相似性，若相似度较低，则拾起该数据并随机移动。拾起概率和放下概率分别如下：

$$P_{pick}(u)=\left[\frac{k_p}{k_p+s(u,v)}\right]^2 \tag{11}$$

$$P_{drop}(u)=\begin{cases}2\times s(u,v) & if \quad s(u,v)<k_d\\ 1 & if \quad s(u,v)>k_d\end{cases} \tag{12}$$

其中：P_{pick},P_{drop} 分别为拾起概率和放下概率，k_p,k_d 为常数，$s(u,v)$ 相似度计算公式如下：

$$s(u,v)=\begin{cases}\frac{1}{S^2}\sum_{u\in Neigh(v)}\left[1-\frac{d(u,v)}{\alpha}\right] & if \quad s(u)>0\\ 0 & otherwise\end{cases} \tag{13}$$

其中：S 为蚂蚁的可见度，$d(u,v)$ 为节点 u,v 之间的距离，α 为相异度因子。

在 LF 算法中，P_{pick},P_{drop} 分别与生成的随机数进行比较，进而执行相应的操作，这种随机方式会使得一个数据被反复地拾起放下，导致聚类速度变慢。同时由于随机性，会导致原本相似的对象被拾起，原本不相似的对象反而被放下，进而导致聚类结果的准确率下降。此外，由于相似性度量和蚁群搜索方式不适用于 PPI 网络，进而导致召回率不高，针对以上问题，本文提出 FGCDACC-DPC 算法。

2.2　FGCDACC-DPC 算法描述

针对蚁群算法应用于静态 PPI 网络聚类问题中，收敛速度慢，聚类结果具有较大随机性以及召回率和准确率不高等问题，为提高算法效率，提出一种基于动态加权 PPI 网络复合物挖掘算法 FGCDACC-DPC。将该算法应用于构建的动态加权网络，能够保证识别的复合物真实可靠。该算法基于 PPI 网络拓扑结构直接对节点聚类，主要分为三个部分：复合核的构建、基于模糊粒度和紧密度的蚁群聚类以及基于局部和全局的权值更新。

2.2.1 复合核的构建

针对 LF 蚁群算法中蚂蚁位置的随机生成会降低算法效率的问题，为加快算法的收敛速度，在 FGCDACC-DPC 算法中蚂蚁的初始位置不是随机在二维平面上生成，而是随机地从一个复合核 C 出发，这样选取初始位置能够在一定程度上提高聚类效率，又可避免陷入局部最优，且复合核的构建为扩充为复合物奠定基础。复合核的构建思想如下：a）由于蛋白质的关键性是复合物的内在特性，关键蛋白质往往大量集中在某些复合物中[25]，因此本文选取每个时刻子网中所有关键蛋白质作为种子节点集合;b）由于复合物是一个功能相似且高度共表达的稠密子图，因此需要判断构造的复合核是否满足相互作用阈值、密度阈值和连续共表达次数的条件。本文基于以上两个特征来构建复合核，其中复合核的密度计算公式如下：

$$den(C)=\frac{2m}{n(n-1)} \tag{14}$$

其中：m 表示复合核 C 的边数，n 表示复合核 C 的节点数。

复合核的构建过程如下：

输入： T_i 时刻下的瞬态加权 PPI 网络 DWG_i 和关键蛋白质，相互作用阈值 η，密度阈值 d ,连续共表达次数 m

输出： T_i 时刻下的复合核 $\{C_1,C_2...C_k\}$

a)根据式（6）计算每个关键蛋白质节点的 $SoCE_{cc}$ 值，按降序排列放入有序队列 Q_1。

b)从 Q_1 中取 $SoCE_{cc}$ 值最大的节点初始化一个复合核 C，将满足 η 并且连续共表达次数大于等于 m 的直接邻居节点加入复合核 C。

c)计算复合核 C 是否满足密度阈值 d，满足转到步骤 4;不满足，递归删除 $SoCE_{cc}$ 值小的节点直至满足条件。

d)得到复合核 C，存入结果队列 Q_2 中，从有序队列 Q_1 中删除复合核 C 中全部的节点。

e)重复步骤 b)~d)，直到有序队列 Q_1 为空。

2.2.2 基于模糊粒度和紧密度的蚁群聚类

为提高算法性能，本文采用模糊粒度和紧密度对拾起放下规则进行改进，而不是基于拾起放下概率与随机数的结果进行聚类，有效降低算法的随机性。其中以模糊粒度作为拾起规则，一方面减少参数的设置，降低计算复杂度，提高聚类速度；另一方面相似度函数能更准确地反映蛋白质与复合核之间的紧密程度，提高聚类准确性。以紧密度作为放下规则，能够对形成的初始聚类结果进行修正，提高聚类效果。

定义 3　模糊粒度[26]。设给定论域 R，ε_A 是 R 到闭区间[0,1]的任一映射（可表示为 ε_A: R→[0,1]），如果有 $r\rightarrow\varepsilon_A(r)$，则 r 为论域 R 的一个模糊子集 A，$\varepsilon_A(r)$ 为 r 对此模糊子集 A 的隶属度。

要衡量复合核内节点 v 与其邻域节点 u 是否相似，首先计算复合核 C 与邻域蛋白质 u 之间的 CWM 值之和，再取其均值作为论域 R，相似度函数可表示为 R 上的一个模糊子集 A，因此基于模糊粒度的相似性函数可表示为：

$$\varepsilon_A(u)=\begin{cases}1-\dfrac{\alpha}{\frac{1}{|C|}\sum_{v\in C,u\in Neigh(C)}CWM(u,v)} & \frac{1}{|C|}\sum_{v\in C,u\in Neigh(C)}CWM(u,v)\geq\alpha\\ 0 & else\end{cases} \tag{15}$$

其中：$|C|$ 为复合核内的节点数，α 为相异度因子，α 取值应该尽可能合理，α 太大，会生成许多稀疏的小类，将直接导致原本能聚到同一个簇的节点不能聚集到同一个簇，反之，会导致原本属于两个簇的节点被划分到同一个簇中。

采用 ε_A 作为衡量是否拾起的标准，如果 ε_A 大于初始粒度 P，则说明该节点 u 与复合核的相似度较大，则对其进行拾起操作，反之不对其进行操作。初始粒度 P 对聚类结果有直接的影响，初始粒度 P 越大，能够满足阈值条件的相互作用就越少，生成的模块数量就多；反之，P 越小，聚类数目越多；粒度 P 的取值在实验 3.3 部分做具体阐述。

定义 4　紧密度 [27]是保证形成高内聚复合物的条件之一，蛋白质 u 到一个复合物 PC 的紧密度 $CD(u,PC)$ 定义如下：

$$CD(u,PC)=\frac{\sum_{u,v_1\in PC}d^{in}(u,v_1)}{\sum_{u\in PC,v_2\notin PC}d^{out}(u,v_2)} \tag{16}$$

其中：$d^{in}(u,v_1)$ 表示蛋白质 u 与复合物 PC 内其他蛋白质 v_1 连接边的权值；$d^{out}(u,v_2)$ 表示蛋白质 u 与复合物 PC 外其他蛋白质 v_2 连接边的权值。

FGCDACC-DPC 算法采用式（15）（16）和（9）代替式（11）～（13）作为拾起放下规则，并调整拾起放下规则。大致思想为：每只蚂蚁的职责是遍历复合核邻域内所有未访

问的节点，并且能装载数据[12]。蚂蚁在 PPI 网络上移动，通过基于模糊粒度的拾起规则不断装载数据形成自身的聚类结果（解），每只蚂蚁对应一个可能解。在形成聚类的过程中，每个复合核初始化一个簇，蚂蚁随机选择一个复合核，搜索复合核邻域内的节点，当蚂蚁遍历完当前复合核邻域内所有满足条件的节点或者达到最大装载量时，蚂蚁随机选择下一个复合核开始下一轮搜索。重复上述过程，直到所有复合核均被蚂蚁遍历完，得到初始聚类结果。根据紧密度对初始聚类结果进行修正，舍弃一些外部连接紧密，内部连接松散的节点。蚁群聚类过程如下所示。

输入： T_i 时刻下的瞬态加权 PPI 网络 DWG_i 和复合核 $\{C_1,C_2...C_k\}$，蚂蚁个数 Num，最大装载量 $L_{\max}$，初始粒度 P

输出： T_i 时刻下的复合物集合 CS

a)在结果队列 Q_2 中随机选择一个复合核 C 作为蚂蚁的初始位置。

b)根据式（15）计算蚂蚁邻域范围内（直接邻居）节点 u 的 ε_A，将满足条件的邻居节点拾起，并前进到该节点，更新复合核和蚂蚁邻域范围；若无满足条件的节点，转到步骤 d)，否则转到步骤 c)。

c)判断蚂蚁装载量（标准复合物的最大规模）是否达到最大，若未达到最大装载量，重复步骤 b)，继续对蚂蚁的新邻域内节点进行聚类；否则转到步骤 d)。

d)得到复合核 C 对应的初始结果，从结果队列 Q_2 中删除复合核 C。判断结果队列 Q_2 是否为空，若不为空，随机选择一个复合核作为蚂蚁的初始位置，返回步骤 b)；否则转到步骤 e)。

e)根据式（16）计算节点 u 与复合物 PC 的紧密度，将紧密度小于 1 的节点舍去，得到复合物 PC。输出复合物集合 CS。

聚类完成之后选取模块性 M [28]作为评价解的质量好坏的指标。M 值越大，说明解的质量越好，M 函数的定义如下：

$$M=\sum_{PC=1}^{CS}\left[\frac{e_{PC}}{|E|}-\left(\frac{d_{PC}}{2|E|}\right)^2\right] \tag{17}$$

其中：CS 为预测到的复合物的个数，e_{PC} 是复合物 PC 内部节点之间连接数之和，d_{PC} 是复合物 PC 内部节点度之和，$|E|$ 表示在整个 PPI 网络所有连接数之和。

2.2.3 局部和全局权值更新策略

针对蚁群算法中聚类准确性不高的问题，提出一种局部更新策略。该方法采用一种功能信息传递机制，通过不同代蚁群之间的信息传递，将上一次迭代的最优解信息通过权值进行传递，使相似的数据在下一次迭代中被分配到同一簇的概率增大，同时减小不相似数据被分配到同一簇的概率。通过不断迭代，使得聚类结果趋于最优。局部权值更新公式如下所示：

$$CWM(u,v)=(1+PC_{uv})CWM(u,v) \tag{18}$$

其中：PC_{uv} 表示在上次迭代的最优解中，蛋白质 u,v 共享复合物的概率，以此作为一种增强系数，公式如下所示：

$$PC_{uv}=\frac{|C_u\cap C_v|^2}{|C_u||C_v|} \tag{19}$$

其中：C_u,C_v 分别为蛋白质 u,v 所属复合物的集合，$C_u\cap C_v$ 表示同时包含蛋白质 u,v 的复合物集合。

研究表明，连续时刻的复合物之间具有一定的相关性[29]，本文结合 PPI 时序网络的功能相关性和蚁群算法的信息传递机制，提出一种全局权值更新策略。大致思想为：a）假设在 T_{i-1} 时刻网络中活性蛋白质 u,v 出现在同一复合物的次数越多，说明该对蛋白质 u,v 功能越相似，那么在 T_i 时刻网络中，如果蛋白质 u,v 仍具有活性，那么该对蛋白在 T_i 网络出现在同一复合物中的概率要比在 T_{i-1} 时刻不属于同一复合物的蛋白对的概率更大;b）假设蛋白质 u,v 在 G_{i-1} 和 G_i 网络上都具有连续活性并且有相互作用，说明该条相互作用是可靠的和稳定的，赋予该相互作用一个高权值 β，假设蛋白质 u,v 只在 G_i 网络上都具有活性和相互作用，说明该条相互作用是比较可靠，赋予该相互作用一个相对较低的权重 ∂。基于以上两点，设计了全局权值更新策略，该策略基于 PPI 网络的时序性，将上一时刻网络的聚类结果通过 CWM 的正反馈传递给下一时刻的网络，有效增加属于同一簇的两个蛋白质之间的相互作用程度，加快收敛速度。全局更新公式如下所示：

$$CWM^i(u,v)=\begin{cases}(1+\partial)^{n_{uv}^{i-1}}CWM^i(u,v), & G_{uv}^{i-1}=0 \quad or \quad G_{uv}^{i}=1\\(1+\beta)^{n_{uv}^{i-1}}CWM^i(u,v), & G_{uv}^{i-1}=1 \quad and \quad G_{uv}^{i}=1\\0, & otherelse\end{cases} \tag{20}$$

其中：n_{uv}^{i-1} 和 n_{uv}^{i} 表示在 T_{i-1} 和 T_i 时刻瞬时网络的最优解中，蛋白质 u,v 出现在同一复合物中的次数，$0\le\alpha<\beta\le 1$，∂ 和 β 为常数。在实验中，分别设置 ∂ 和 β 为 0.1 和 0.2。

2.3 算法步骤

FGCDACC-DPC 算法步骤如下所示：

输入： 各个时刻下的动态加权 PPI 网络 DWG_i

输出： 各个时刻下的复合物

a)初始化参数：相互作用阈值 η，密度阈值 d，连续共表达次数 m，相异度 α，初始粒度 P，最大迭代次数 N，蚂蚁个数 Num，最大装载量 $L_{\max}$，∂，β

b)挖掘各个时刻下的复合物

```
for i=1 to 12 do
  调用复合核构建方法
  for iter=1 to N
    for k=1 to Num
      调用蚁群聚类方法
    end for
    根据式（17）计算蚁群的最优解
    根据最优解和式（18）对相互作用权值进行局部更新
  end for
  根据第 T_i 时刻的最优解和式（20），全局更新第 T_{i+1} 时刻的相互作用权值
end for
输出不同时刻 T_i 的蛋白质复合物
```

2.4 算法的时间复杂度

FGCDACC-DPC 算法的时间复杂度主要由三部分构成。初始化参数的时间复杂度为 O(1)。构建复合核的时间复杂度为 $O(m_1|n_i|d_{\max}l)$。蚁群聚类的时间复杂度，在最好情况下，除复合核的剩余节点 $m_2|n_i|$ 在每一次聚类中至少被访问一次，聚类的时间复杂度 $O(N{*}Num{*}m_2|n|l)$；在最坏情况下，剩余节点 $m_2|n_i|$ 在每一次聚类中都被访问却又没有拾起，此时每个节点被访问了 $m_1|n_i|$ 次，聚类的时间复杂度 $O(N{*}Num{*}m_1m_2|n_i|^2 l)$。其中 $m_1|n_i|$ 为 DWG_i 上复合核的数量，$d_{\max}$ 为节点的最大度，l 为动态子网的个数，N 为迭代次数，Num 为蚂蚁数量。由于 $n_i<n$（n 为总节点数），并且 N,Num,l 和 m_1,m_2 均为常量，因此 FGCDACC-DPC 算法性能较好。

2.5 算法后处理

重叠得分[30]通常用来评价检测到的复合物 F_u 与标准库复合物 S_v 的匹配度，定义如下：

$$OS(F_u,S_v)=\frac{|F_u\cap S_v|^2}{|F_u|\times|S_v|} \tag{21}$$

若 $OS(F_u,S_v)\ge t$，则表示预测复合物与标准复合物匹配，t 一般取值为 0.2[30]，OS 值越大说明匹配率越高。采用 FGCDACC-DPC 算法聚类得到结果后，根据式（21）计算预测得到的复合物与标准复合物的重叠得分，当重叠率低于 0.2，则将该复合物删除，重复该过程，得到最终的复合物。

3　实验结果与分析

3.1　数据来源

为验证算法的有效性，本文选用基因表达数据，GO 功能注释数据等相关数据集都相对比较完善的酵母 PPI 网络数据。实验部分所使用到的几种数据如下所示:

a)酵母 PPI 网络来自 DIP 数据库[31]（2010 年 10 月 10 日的版本），经去除重操作后，该数据库包含 5093 个蛋白质和 24734 组相互作用。

b) GO 功能注释信息下载自基因本体库[32]。

c) 基因表达谱数据选取编号为 GSE3431 的数据[33]，包括 36 个时刻下的 6777 个基因的表达值。经过预处理后，在酵母 PPI 网络中的只有 4981 个基因。本文将没有基因表达数据的蛋白质的基因值设置为 0。根据基因表达谱数据对 PPI 网络预处理后，得到 12 个时刻瞬态子网的活性蛋白质数目及其相互作用数目，具体数据如表 1 所示。

采用 CYC2008 [34]作为标准复合物数据集。其中包含 408 个标准复合物，簇的最大规模为 81，考虑可扩展性，因此本文将最大装载量 $L_{\max}$ 设置为 90。

d) 关键蛋白质数据通过整合 MIPS[35], SGD[36], DEG[37], SGDP[38]4 个数据库中的数据得到，其中包含 1285 个关键蛋白质，只有 1167 个关键蛋白在酵母 PPI 网络中。

表 1　瞬态 ppi 网络的活性蛋白质及其相互作用数目

Table 1　Number of active proteins and interactions in each dynamic PPI networks

时间点	1	2	3	4	5	6
活性蛋白质数目	1638	1742	1659	1444	1368	1211
相互作用数目	7574	8497	8262	6697	6250	5264
时间点	7	8	9	10	11	12
活性蛋白质数目	1221	1444	1756	1285	1410	1249
相互作用数目	5438	7109	8698	5999	6598	5306

3.2　评价指标

实验采用正确率，查全率和 $F-measure$ 来对算法聚类效果进行评估[30]，其计算公式如下：

$$precision(F,S)=\frac{|F\cap S|}{|F|} \tag{22}$$

$$recall(F,S)=\frac{|F\cap S|}{|S|} \tag{23}$$

其中：F 表示预测的蛋白质复合物，S 表示标准库中的蛋白质复合物。

综合考虑正确率和查全率对聚类结果的影响，采用 $F-measure$ 综合度量模块的聚类结果。

$$F-measure=\frac{2\times precision\times recall}{precision+recall} \tag{24}$$

3.3　参数分析

FGCDACC-DPC 算法中，初始粒度 P 需要用户自定义，并且直接影响聚类数目。因此为使得初始粒度 P 的取值较为合理，本文在保证其他参数相同的情况下，独立运行 20 次，取 20 次结果的平均值进行分析。实验中使用的参数设置如下：$\eta=0.35$，$m=5$，$\alpha=0.35$，$N=20$，$d=0.7$，Num=15。

图 2 展示了粒度 P 在不同取值下精度、召回率以及 $F-measure$ 值相应的变化情况。从图 2 中可以看到，结果的精度随着粒度 P 值的增大而增大，由于 P 值的增大，满足条件的相互作用变少，能够避免和某一复合物相似度较低的节点的加入，从而检索到更少的无用复合物，所以整体的正确率呈上升趋势；与此同时，召回率会随着 P 值的增加而下降，因为要求更严格，能够匹配的功能模块的数目变少。P 值从 0 增加到 0.55，$F-measure$ 值一直处于上升趋势，到达 0.55 之后，$F-measure$ 值逐渐趋于平稳。因此本文将 P 设置为 0.55 较为合理。

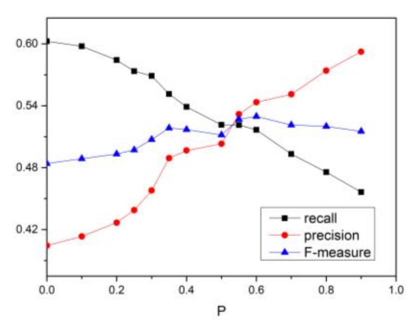


图 2　粒度 P 值与评价指标的关系

Fig. 2　Performance of FGCDACC-DPC with different granularity p

3.4　综合性权值度量有效性分析

为验证综合性权值度量 CWM 的有效性，实验使用 FGCDACC-DPC 算法，分别以不同的加权方式对动态 PPI 网络加权，比较在不同重叠率阈值下，被识别的已知复合物的比例。将文献[16]中的加权方法 W 与 CWM 加权进行对比实验，图 3 为重叠率阈值在[0.2,1.0]内变化时，两种加权策略所检测到的已知蛋白质复合物的比例。

从图 3 中可以看出本文加权策略的检测结果明显优于文献[16]的加权策略，尤其在重叠率阈值为 0.3 时，FGCDACC-DPC 算法使用 CWM 加权识别的蛋白质复合物比例要比使用 W 加权识别的比例高 20.8%。由于文献[16]中 W 加权只整合了边聚集系数和持续共表达的长度，因此只能反映 PPI 网络的拓扑特性和时序性，考虑的比较单一。而 CWM 加权综合考虑了蛋白质相互作用网络的拓扑特性和生物特性，并且利用 GO 注释信息和基因表达数据为网络加权，有效减少假阴性和假阳性带来的负面影响，因此基于 CWM 加权的动态 PPI 网络蛋白质复合物的预测效果较好。

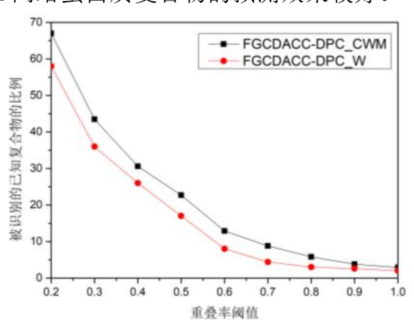
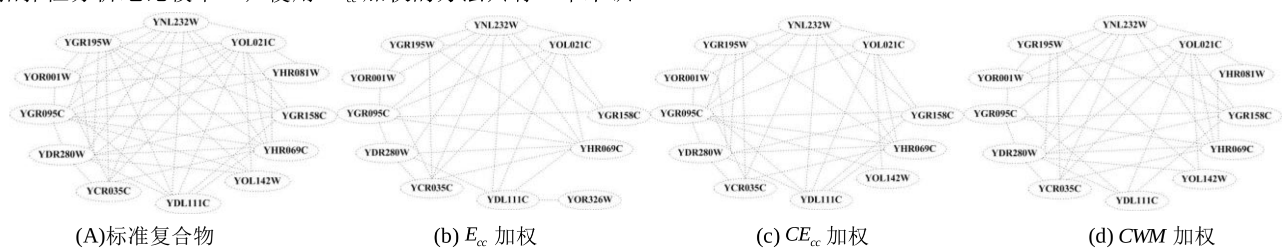


图 3　不同加权策略挖掘的复合物对比结果

Fig. 3　Comparison of complexes mined by different weighting strategies

为进一步检验综合性权值的有效性，分别使用不同加权方法与 *CWM* 加权进行对比实验。图 4 是采用不同加权方法检测到的蛋白质复合物与标准数据库的对比结果。从图 4 中可以看到，使用 *Ecc* 加权的方法未识别 YHR081W 和 YOL142W 两个蛋白质，而且 YDL111C 节点错误挖掘了一个 YOR326W 节点，是因为边聚集系数只考虑节点间边的紧密度，对网络拓扑性分析地比较单一；使用 CE_{cc} 加权的方法只有一个未识别的蛋白质，因为点边聚集系数不仅考虑了节点间边的关系，还考虑了每个节点的重要性，对蛋白质网络的拓扑性考虑的比较全面，但忽略了蛋白质之间的生物特性；综合性权值度量既考虑网络拓扑性，同时又结合了 GO 注释信息和基因表达数据，对蛋白质网络进行了全面的分析，能够更加贴近真实网络，因此最终的聚类效果较好。



(A)标准复合物　　(b) E_{cc} 加权　　(c) CE_{cc} 加权　　(d) *CWM* 加权

图 4　不同加权方法识别 nuclear exosome complex 复合物

Fig. 4　Nuclear exosome complex detected by different weighting methods

3.5　3.5 权值更新策略有效性分析

为检验权值更新策略对当前时刻复合物的聚类效果，本文分别基于使用权值更新策略的 FGCDACC-DPC 算法和基于未使用权值更新策略的 FGCDACC-DPC 算法，在 12 个子网中进行复合物检测，将每个时刻下两种情况的检测结果进行对比分析。

图 5 为两种情况的 $F-measure$ 对比结果。从图 5 中可以看出，有 2 个时刻两种情况的 $F-measure$ 值持平，有 8 个时刻使用权值更新策略的 FGCDACC-DPC 算法超过了未使用权值更新策略的算法，尤其是在 9、10、11 和 12 时刻 $F-measure$ 值有明显提高。假设前一时刻的聚类比较准确，那么当前时刻根据前一时刻的聚类结果进行权值更新，可以保证当前时刻的相互作用更真实可靠，进而提高聚类效果。

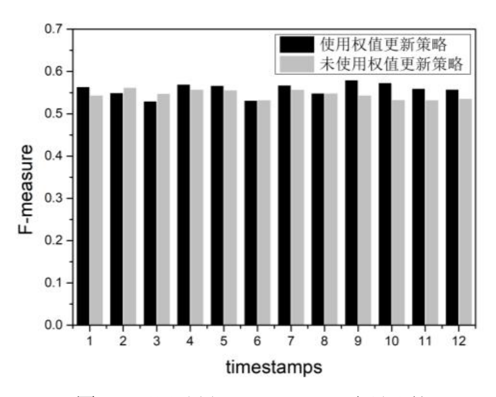


图 5　不同时刻下 $F-measure$ 度量比较

Fig. 5　The value of f-measure with different strategies in different periods of time

图 6 显示 FGCDACC-DPC 算法在 12 个瞬态网络中的 $precision, recall$ 和 $F-measure$ 值。不难看出随着精确度的上升召回率逐渐下降。在第 6 个时刻，精确度值和召回率值分别到达最高点和最低点，在第 7 个时刻之后，精确度、召回率以及 F 度量值逐渐趋于平缓，精度在 11、12 时刻有所上升。

3.6　算法有效性比较

3.6.1 性能分析

为验证 FGCDACC-DPC 算法在动态 PPI 网络的有效性，分别选用传统聚类算法 MCODE[3]，RNSC[4]，新型聚类算法 MCL[6]，COACH[7]以及基于蚁群聚类的算法 JSACO[12]、ACC-FDM[10]以及 ACC-DPC[18]等与 FGCDACC-DPC 算法进行对比实验，分析各算法的精度、召回率以及 $F-measure$ 值。图 7 为各算法在三种度量指标上的对比结果。

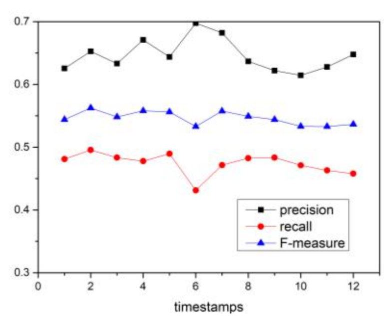


图 6　fgcdaco-dpc 算法各时刻的度量值

Fig. 6　Evaluation value of FGCDACO-DPC algorithm in different periods of time

从图中可以看出，FGCDACC-DPC 算法具有最高的 $F-measure$ 值，分别比 MCODE，MCL，COACH，RNSC，ACC-DPC，JSACO 算法和 ACC-FMD 提高了 144.3%，61.06%，19.24%，37.58%，17.49%,42.161%，25.52%。其主要原因有：一方面构建的动态加权 PPI 网络更加贴近真实的 PPI 网络，降低假阳性和假阴性对聚类准确性的影响；另一方面对拾起放下的改进策略和权值更新策略能够有效提升算法的 $F-measure$ 。该算法在精度上位列第二，仅次于 JSACO 算法，说明构建的动态网络包含较少的假阳性。该算法在召回率上的表现较优，分别比 MCODE，MCL，COACH，RNSC，ACC-DPC，JSACO 算法和 ACC-FMD 提高了 252.2%，38.025%，7.08%，14.01%，27.17%，95.758%和 40.157%。虽然构建的动态网络会缺少一定量的蛋白质，这样可能会导致召回率有所下降，但加权方式的有效性使得网络中含有较少的假阴性，因此召回率整体上提高了。综合衡量三个指标值，该算法性能较优。

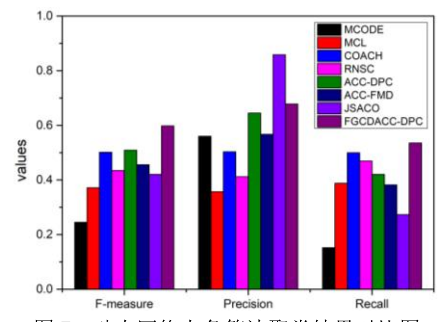


图 7　动态网络上各算法聚类结果对比图

Fig. 7　Comparison results of different algorithms in dynamic PPI network

为进一步评估 FGCDACC-DPC 算法的聚类性能，分别从各类算法识别的复合物的个数、簇平均大小、覆盖蛋白质数以及运行时间四方面进行分析。从表 2 中可以看出，FGCDACC-DPC 算法识别复合物的平均大小和覆盖蛋白质数比其他算法识别的结果都要更加接近标准类；虽然识别的复合物个数为 637，仅次于 MCL 算法，但 MCL 算法覆盖的蛋白质却有 4096 个，其准确率比 FGCDACC-DPC 算法低。

为验证该算法的时间效率，将 FGCDACC-DPC 算法与各种基于蚁群聚类的算法进行对比实验。从表 2 中可以看出本文算法时间性能较优，首先是因为该算法是基于小规模动态加权 PPI 网络聚类的，克服了蚁群算法应用于大规模 PPI 网络收敛速度慢的问题；其次改进的拾起放下规则和权值更新的有效性，能够有效减少计算量和访问却不拾起的次数，进而缩短聚类时间。因此该算法比 ACC-DPC 和 ACC-FMD 算法的时间效率要高。虽然 FGCDACC-DPC 算法的运行时间稍次于 JSACO 算法，但该算法的其他指标却高于 JSACO 算法。整体上看，FGCDACC-DPC 算法具有良好的性能。

3.6.2 聚类结果分析

这一部分主要分析 FGCDACC-DPC 算法的聚类结果，表 3 为采用该算法识别的其中 6 个蛋白质复合物。通过分析预测复合物中正确和错误的聚类结果来评价该算法的聚类效果。从表 3 可以看出，预测复合物 2、3、5 和 6 与标准复合物为完美匹配，说明采用 FGCDACC-DPC 算法检测的蛋白质复合物与真实蛋白质复合物更加贴近，更具生物意义。

表 2　各种挖掘蛋白质复合物算法的性能比较

Table 2　Performance comparision of various algorithms for mining protein complexes

算法	簇的个数	平均大小	覆盖蛋白质	运行时间（/s）
标准类	408	4.71	1628	--
MCODE	107	6.5	1299	--
MCL	623	6.57	4096	--
COACH	903	8.9	1999	--
RNSC	160	3.89	1489	--
ACC-DPC	237	7.8	1785	1524
ACC-FMD	283	9.5	1832	9133
JSACO	665	3.627	1958	671
FGCDACO-DPC	637	4.98	1921	706

表 3　fgcdacc-dpc 算法识别的 6 个复合物的结果分析

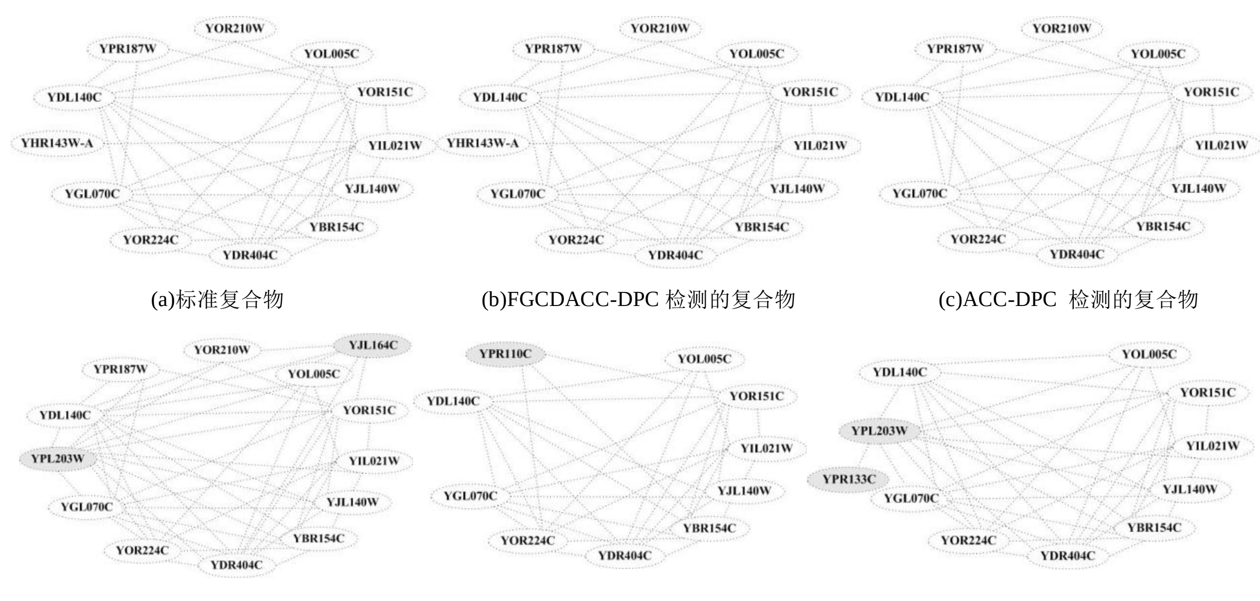
Table 3　Analysis of six protein complexes detected by FGCDACC-DPC algorithm

聚类序号	复合物名称	标准复合物	正确聚类的蛋白质	错误聚类的蛋白质	*OS*
1	Exocyst complex	YJL085W YBR102C YLR166C YGL233W YER008C YDR166C YIL068C YPR055W	YBR102C YLR166C YGL233W YER008C YDR166C YIL068C YPR055W		0.875
2	Cbf1p/Met4p/Met28p complex	YJR060W YIR017C YNL103W	YJR060W YIR017C YNL103W		1
3	TRAMP complex (Air2p)	YDL175C YJL050W YOL115W	YDL175C YJL050W YOL115W		1
4	histone deacetylase complex	YGL194C YIL112W YDR155C YOL068C YKR029C YBR103W YCR033W	YGL194C YIL112W YDR155C YOL068C YKR029C YBR103W YCR033W	YMR173C	0.875
5	cAMP-dependent protein kinase	YIL033C YJL164C YPL203W YKL166C	YIL033C YJL164C YPL203W YKL166C		1
6	DNA-directed RNA polymerase II complex	YOR210W YOL005C YOR151C YIL021W YJL140W YBR154C YDR404C YOR224C YDL140C YPR187W YGL070C YHR143W-A	YOR210W YOL005C YOR151C YIL021W YJL140W YBR154C YDR404C YOR224C YDL140C YPR187W YGL070C YHR143W-A		1

为更加直观地分析聚类结果，本文将 DNA-directed RNA polymerase II complex 复合物的检测结果进行可视化。图 8 展示的是不同算法检测该复合物的预测结果，其中灰色节点为聚类错误的蛋白质。（a）是标准复合物；（b）是 FGCDACC-DPC 算法的检测结果，正确检测该复合物的全部蛋白质；（c）是 ACC-DPC 算法的检测结果，正确检测到 11 个蛋白质，只有蛋白质 YHR143W-A 未被检测出来，是因为该节点只与簇内 YIL021W 相连，并且与簇外连接更加紧密；（d）是 ACC-FMD 算法的检测结果，检测到 10 个蛋白质，错误检测两个非复合物内蛋白质，其中蛋白质 YPL203W 错误替代 YHR143W-A，只因为 YPL203W 与簇内所有蛋白质都连接紧密。从图 8（c）和（d）的聚类结果中可以看出，在使用同种算法的情况下基于动态网络挖掘的复合物更加准确；（e）和（f）为 MCL 和 MCODE 算法的检测果，这两种算法都只正确检测到 9 个蛋白质，其中 MCL 算法检测结果中的蛋白质 YPR110C 错误替换 YPR187W，MCODE 算法错误检测两个蛋白质。综述所述，基于动态加权 PPI 的 FGCDACC-DPC 算法的检测结果更加接近标准复合物，进一步说明该算法的有效性。

4　结束语

本文面临的首要问题是如何有效减少 PPI 网络中的假阳性和假阴性数据，进而构建真实可靠的动态加权网络。并且针对基于拾起放下规则的蚁群聚类算法中的缺陷，本文该如何根据 PPI 网络的拓扑特性设计出一种有效的相似性函数，以准确描述节点与复合核的紧密程度，如何根据蛋白质复合物的结构特征设计出一种较优的扩张方法，以优化搜索过程、提高召回率和聚类速度，以及如何根据蚁群信息传递机制和时序网络特性提出一种策略来传递最优解信息，以提高聚类准确性。针对以上问题，本文首先基于静态 PPI 网络的拓扑特性，结合基因表达数据和 GO 注释信息，构建更加可靠的动态加权 PPI 网络，并提出一种基于蚁群聚类的动态 PPI 网络蛋白质复合物挖掘算法 FGCDACC-DPC。与其他蚁群聚类算法相比，该算法充分利用蛋白质的关键性和复合物的形成机制来构建更加贴近真实复合物核心的复合核，以作为扩张的基础，然后将模糊粒度相似度和紧密度的概念应用于聚类中，并对拾起放下规则进行改进，以降低算法随机性和计算复杂度以及有效提高聚类速度和召回率，并且通过不断更新局部和全局权值以传递最优解信息，大大提升算法准确率。实验结果验证了动态加权 PPI 网络的有效性。结果也表明，相比于 MCODE，RNSC，MCL，COACH，JSACO、ACC-FDM 以及 ACC-DPC，FGCDACC-DPC 算法具有较强的蛋白质复合物检测能力，挖掘的蛋白质复合物既满足拓扑结构上的稠密性，又更加贴近生物意义上的复合物，并且在三种评价指标上都取得较好的结果。



(a)标准复合物 (b)FGCDACC-DPC 检测的复合物 (c)ACC-DPC 检测的复合物

(d)ACC-FMD 检测的复合物 (e)MCL 检测的复合物 (f)MCODE 检测的复合物

图 8 不同算法检测的 DNA-directed RNA polymerase II complex 复合物结果对比

Fig. 8 DNA-directed RNA polymerase II complex detected by different algorithms

虽然该算法在准确率和召回率上有所提高，收敛速度过慢的问题得到些许改善，但时间性能未得到显著提升；同时由于连续时刻的复合物之间具有一定的相关性，识别的蛋白质复合物之间重叠过多，这些问题还有待进一步研究。目前本文主要基于酵母 PPI 网络，未来的研究重心应放在人类蛋白质和致病基因预测上，此外可基于检测的蛋白质复合物，根据复合物中大部分功能已知的蛋白质来预测那些功能未知的蛋白质，以上这些将是下一步的研究工作。

参考文献：

[1] Ji Junzhong, Zhang Aidong, Liu Chunnian, *et al.* Survey: functional module detection from protein-protein interaction networks [J]. IEEE Trans on Knowledge & Data Engineering, 2014, 26(2): 261-277.

[2] Zhao Bihai, Wang Jianxin, Li Min, *et al.* A new method for predicting protein functions from dynamic weighted interactome Networks [J]. IEEE Trans on Nanobioscience, 2016, 15(2): 131-139.

[3] Bader G D, Hogue C W V. An automated method for finding molecular complexes in large protein interaction networks [J]. BMC Bioinformatic, 2003, 4(1): 20-29.

[4] King A D, Przulj N, Jurisica I, *et al.* Protein complex prediction via cost-based clustering [J]. Bioinformatics, 2004, 20(17): 3013-3020.

[5] Aldecoa R. Jerarca: Efficient analysis of complex networks using hierarchical clustering [J]. PLoS One, 2010, 5(7): e11585.

[6] 雷秀娟，高银，郭玲，等. 基于拓扑势加权的动态 PPI 网络复合物挖掘方法 [J]. 电子学报，2018, 46(1): 145-151. (Lei Xiujuan, Gao Yin, Guo Ling, *et al.* Mining protein complexes based on topology potential weight in dynamic protein-protein interaction networks[J]. Acta Electronica Sinica, 2018, 46(1):145-151)

[7] Lei Xiujuan, Liang Jing. Neighbor Affinity-Based Core-Attachment Method to Detect Protein Complexes in Dynamic PPI Networks [J]. Molecules, 2018, 22(7): 101-112.

[8] 蔡磊鑫，陈荣，吕强，等. 基于谱聚类的蛋白质对接复合物最佳预测结构的挑选方法 [J]. 小型微型计算机系统，2015, 36(10): 2365-2368. (Cai Leixin, Chen Rong, Lyu Qiang, *et al.* Selecting the best of the predictions of compound structures for protein docking by spectral clustering [J]. Journal of Chinese Computer Systems, 2015, 36(10): 2365-2368.)

[9] Ji Junzhong, Liu Zhijun, Zhang Aidong, *et al.* Improved ant colony optimization for detecting functional modules in protein-protein interaction networks [J]. Communications in Computer & Information Science, 2012, 308(2): 404-413.

[10] Ji Junzhong, Liu Hongxin, Zhang Aidong, *et al.* ACC-FMD: ant colony clustering for functional module detection in protein-protein interaction networks [J]. International Journal of Data Mining & Bioinformatics, 2015, 11(3): 331-363.

[11] 冀俊忠，吕嘉伟. 多粒度 PPI 网络描述模型及其蚁群优化的功能模块检测方法 [J]. 中国科技论文，2014, 9(7):762-769. (Ji Junzhong, Lyu Jiawei. A multiple grain representation model of PPI networks and its ant colony optimization method for functional module detection [J]. China Sciencepaper, 2014, 9(7): 762-769.)

[12] 雷秀娟，黄旭，吴爽，等. 基于连接强度的 PPI 网络蚁群优化聚类算法 [J]. 电子学报，2012, 40(4): 695-702. (Lei Xiujuan, Huan Xu, Wu Shuang, *et al.* Joint strength based ant colony optimization clustering algorithm for PPI networks [J]. Acta Electronica Sinica,2012, 40(4): 695-702.)

[13] Lei Xiujuan, Wang Fei, Wu Fangxiang, *et al.* Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks [J]. Information Sciences, 2016, 329 (6): 303-316.

[14] Tang Xiwei. A comparison of the functional modules identified from time course and static PPI network data [J]. BMC Bioinformatics, 2011, 12(1): 339.

[15] 胡赛，熊慧军，赵碧海，等. 动态加权蛋白质相互作用网络构建及其应用研究 [J]. 自动化学报，2015, 41(11): 1893-1900. (Hu Sai, Xiong Huijun, Zhao Bihai, *et al.* Construction of dynamic-weighted protein interactome network and its application [J]. Acta Automatica Sinica, 2015, 41 (11): 1893-1900.)

[16] Su Lingtao, Liu Guixia, Han Wang, *et al.* GECluster: a novel protein complex prediction method[J]. Biotechnology & Biotechnological Equipment, 2014, 28(4): 753-761.

[17] Shen Xianjun, Yi Li, Jiang Xingpeng, *et al.* Identifying protein complex by integrating characteristic of core-attachment into dynamic PPI

network[J]. PLoS One, 2017, 12 (10): e0186134.

[18] 赵学武，程新党，吕嘉伟，等. 融合时序保持特征和蚁群聚类的动态 PPI 网络复合物识别[J]. 小型微型计算机系统，2017,38(6): 1311-1316. (Zhao Xuewu, Cheng Xindang, Lyu jiawei, *et al.* Identify protein complexes by integrating temporal function continue feature and ant colony clustering on dynamic PPI Networks [J]. Journal of Chinese Computer Systems, 2017, 38(6): 1311-1316.)

[19] Li Min, Yang Jie, Wu Fangxiang, *et al.* DyNetViewer: a Cytoscape App for dynamic network construction, analysis and visualization [J]. Bioinformatics, 2018, 34(9): 1597-1599.

[20] Wang Jianxin, Peng Xiaoqing, Wu Fangxiang, *et al.* Dynamic protein interaction network construction and applications [J]. Proteomics, 2014, 14 (4-5): 338-352.

[21] Wang Jianxin, Li Min, Wang Huan, *et al.* Identification of essential proteins based on edge clustering coefficient [J]. IEEE/ACM Trans on Computational Biology & Bioinformatics, 2012, 9(4): 1070-1080.

[22] 高杨，张燕平，钱付兰，等. 结合节点度和节点聚类系数的链路预测算法[J].小型微型计算机系统,2017, 38(7):1436-1441. (Gao Yang, Zhang Yanping, Qian Fulan, *et al.* Combined with node degree and node clustering coefficient of link prediction algorithm [J]. Journal of Chinese Computer Systems, 2017, 38(7): 1436-1441.)

[23] 赵碧海，李学勇，胡赛，等. 基于关键功能模块挖掘的蛋白质功能预测 [J]. 自动化学报, 2018, 44(1): 183-192. (Zhao Bihai, Li Xueyong, Hu Sai, *et al.* Prediction of Protein functions based on essential functional modules mining [J]. Acta Automatica Sinica, 2018, 44 (1): 183-192.)

[24] Lumer E, Faieta B. Diversity and adaptation in population of clustering ants [C]//Proc of the 3rd International Conference on Simulation of Adaptive Behavior: From Animal to Animals. Cambridge MA: MIT Press, 1994: 499-508.

[25] Hart G T, Lee I, Marcotte E M. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality [J]. BMC Bioinformatics, 2007, 8 (1): 1-11.

[26] Liu Hongbing, Xiong Shengwu, Wu Changan. Hyperspherical granular computing classification algorithm based on fuzzy lattices [J]. Mathematical & Computer Modelling, 2013,57(3-4): 661-670.

[27] 丁玉连,雷秀娟,代才,等.模拟鸽子优化过程的蛋白质复合物识别算法[J].计算机科学与探索,2017,11(8):1279-1287. (Ding Yulian, Lei Xiujuan, Dai Cai, *et al.* Identification of protein complexes by simulating process of pigeon-inspired optimization [J]. Journal of Frontiers of Computer Science and Technology,2017,11(8): 1279-1287.)

[28] 郭茂祖,代启国，徐立秋，等. 一种蛋白质复合体模块度函数及其识别算法 [J]. 计算机研究与发展,2014, 51(10):2178-2186. (Guo Maozu, Dai Qiguo, Xu Liqiu, *et al.* On protein complexes identifying algorithm based on the novel modularity function[J]. Journal of Computer Research and development, 2014, 51(10): 2178-2186.)

[29] Du Nan, Zhang Yuan, Li Kang, *et al.* Evolutionary analysis of functional modules in dynamic ppi networks [C]//Proc of ACM Conference on Bioinformatics. New York: ACM Press, 2012: 250-257.

[30] Wang Jianxin, Peng Xiaoqing, Li Min, *et al.* Construction and application of dynamic protein interaction network based on time course gene expression data[J]. Proteomics, 2013, 13(2): 301–312.

[31] Zhao Bihai, Wang Jianxin, Li Min, *et al.* Prediction of essential proteins based on overlapping essential modules [J]. IEEE Trans on Nanobioscience, 2014, 13(4): 415-424.

[32] Ji Junzhong, Liu Zhijun, Liu Hongxin, *et al.* An Overview of Research on Functional Module Detection for Protein-protein Interaction Networks [J]. Acta Automatica Sinica, 2014, 40(4): 577-593.

[33] Tu Benjamin P, Mcknight Steven L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes[J]. Science, 2005, 310 (5751): 1152-1158.

[34] Zhao Bihai, Wang Jianxin, Li Min, *et al.* Detecting protein complexes based on uncertain graph model[J]. IEEE//ACM Trans on Computational Biology & Bioinformatics, 2014, 11(3): 486-497.

[35] Mewes H W, Frishman D, Mayer K F, *et al.* MIPS: analysis and annotation of proteins from whole genomes [J]. Nucleic Acids Research, 2004, 34(2):169-72.

[36] Cherry J M, Adler C, Ball C, *et al.* SGD: saccharomyces genome database[J]. Nucleic Acids Research, 1998, 26(1): 73-9.

[37] Zhang Ren, Lin Yan. DEG 5. 0, a database of essential genes in both prokaryotes and eukaryotes [J]. Nucleic Acids Research, 2009, 37 (Database issue): D455-D458.

[38] Bruno A, Jef B, Carla C, *et al.* SGDP: Saccharomyces Genome Deletion Project [EB/OL]. http://www-sequence.stanford.edu/group/yeast_deletion_project.